Research Journal of Applied Sciences, Engineering and Technology 10(12): 1441-1444, 2015
DOI: 10.19026/rjaset.10.1846
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: May 1, 2015 Accepted: May 10, 2015 Published: August 25, 2015

Research Article

A Precise Distance Metric for Mixed Data Clustering using Chi-square Statistics

S. Mohanavalli and S.M. Jaisakthi
SSN College of Engineering, Chennai, Tamil Nadu-603110, India

Abstract: In today's scenario, data is available as a mix of numerical and categorical values. Traditional data clustering algorithms perform well for numerical data but produce poor clustering results for mixed data. For better partitioning, the distance metric used should be capable of discriminating the data points with mixed attributes. The distance measure should appropriately balance the categorical distance as well as numerical distance. In this study we have proposed a chi-square based statistical approach to determine the weight of the attributes. This weight vector is used to derive the distance matrix of the mixed dataset. The distance matrix is used to cluster the data points using the traditional clustering algorithms. Experiments have been carried out using the UCI benchmark datasets, heart, credit and vote. Apart from these data sets we have also tested our proposed method using a real-time bank data set. The accuracy of the clustering results obtained are better than those of the existing works.

Keywords: Chi-square statistics, clustering, mixed data attributes

INTRODUCTION

Clustering is a well practiced partitioning algorithm used to separate data objects into similar groups. The quality of a clustering algorithm is dependent on how well the data objects are discriminated (Han *et al*., 2006). The distance metric that is used to measure the similarity between the data objects determines the extent of separation (Ralambondrainy, 1995). The classic clustering algorithms are successful in grouping data objects of same type of features (numerical or categorical). The euclidean distance metric is popularly used for numeric data and chi-square statistics (NIST, 2012) is adopted for categorical attributes. In this era of Big Data, data is available as a mix of numeric and categorical features. In order to properly partition the data, it is necessary to establish a balance between numeric distance and categorical distance. The lack of a single distance metric to measure similarity between data objects with mixed attributes is the motivation for carrying out this research work.

There is a lot of research being carried out to handle mixed data attributes that coexist in data objects. Ahmad and Dey (2007) proposed a new cost function and distance measure based on co-occurrence of values for clustering the categorical data. To obtain the cost, the authors have combined two different cost functions, one for numerical data and another one for categorical data which should be minimized. The weight factor in the cost function is determined automatically from the data points which specifies the significance of the numeric data. The distance is measured using weighted squared value of the Euclidean distance between the data object's attribute value and the value of the attribute at the center of the cluster. Bai *et al*. (2011) proposed a new weighting technique for clustering the categorical data which is the extension of K-mode algorithm. Two weights have been calculated for each attribute in each cluster which is used to identify the subsets of important attributes used for clustering. A weighting k-mode algorithm for subspace clustering of categorical data was proposed by Cao *et al*. (2013). The proposed algorithm presented in this study used complement entropy to automatically compute the weight of all dimensions in each cluster in the k-mode algorithm. The calculated attribute weight is used to identify the subsets of important dimensions that categorize different clusters. A new concept called distribution centroid to represent the prototype of categorical attributes in a cluster was proposed by Ji *et al*. (2013). The calculated distribution centroid and mean were integrated to represent the prototype of a cluster with categorical data and propose a new dissimilarity measure that incorporates the significances of each attribute, to evaluate the dissimilarity between attributes and prototype. A new dissimilarity measure based on the idea of biological and genetic taxonomy was introduced by Cao *et al*. (2012). The authors used a new membership-based dissimilarity measure by taking the distribution of attributes values in the universe and the dissimilarity measure is calculated by improving the Ng's dissimilarity measure. The calculated dissimilarity measure is finally used in the k-mode algorithm for clustering the categorical data. Ji *et al*. (2012) presented a new method to cluster categorical data by integrating

Corresponding Author: S. Mohanavalli, SSN College of Engineering, Chennai, Tamil Nadu-603110, India

mean and fuzzy centroid. Dissimilarity measure is calculated by using the significance of each attribute and distance between categorical values and fuzzy clustering algorithm is applied for clustering the categorical data. He *et al*. (2011) proposed a clustering algorithm which generalized the k-modes clustering algorithm using attribute value weighting in dissimilarity computation.

It is evident from the existing research work that there is still a need for design of a method to measure the similarity between data objects with mixed data attributes. The objective of this research work is to propose a chi-square statistic based weighting scheme to achieve a balance between the numerical and categorical attributes. The clustering results using our proposed distance metric proved to be better than those of the existing methods.

MATERIALS AND METHODS

In reality the field expert has knowledge about the class distribution for few samples. This domain knowledge is used in the chi-square statistic to learn the relevance of the attributes in clustering the data objects.

Materials used in this research work: The experiments were conducted with UCI data sets (UCI ML Dataset, year) namely Heart Disease, Credit Approval and Vote and a real time bank dataset. The proposed clustering method was implemented using R packages (Core Team, 2013).

Heart disease dataset: The heart disease dataset has 303 instances with 14 attributes, both categorical and numerical. The class distribution attribute refers to the presence of heart disease in the patient. The chi-square statistical test was performed to determine the influence of each attribute on the class attribute.

Credit approval dataset: The credit approval dataset consists of 690 instances having 15 attributes, featuring a good mix of continuous and nominal attributes, in the financial domain. This dataset is subject to chi-square statistics, to determine the relevance of each attribute in credit approval of a customer.

Vote dataset: The Vote dataset has 435 instances with 16 categorical attributes. The chi square statistics is again applied here to classify the data records as republic or democratic.

Bank dataset: The real time bank dataset consists of 1021 instances having 17 attributes, with both categorical and numerical data. The class distribution attribute is a binary value indicating whether the customer would prefer a new product of the bank. Chi-square test was applied on a known bunch of customers to learn the weight/relevance of each attribute to a class attribute which was used for clustering of the records.

Table 1: Contingency table for class distribution

Class	Attribute value 1	Attribute value 2	Total
Label 1	a	b	a+b
Label 2	c	d	c+d
Total	a+c	b+d	a+b+c+d = N

Proposed method:

Chi-square statistics based weighting scheme: A chi-square (X^2) statistic is used to investigate whether distributions of categorical variables differ from one another. The chi-square statistic compares the counts of categorical responses between two (or more) independent groups. The chi-square test involves population of contingency table, with the frequency of the attributes (categorical or continuous) considered, together with the frequency of the class attribute (NIST, 2012). Then, we calculate the chi-square distribution value and determine its ratio to the sum of all chi-square values as the weight to be used in the clustering process. The contingency table for an attribute with its categorical values and corresponding label distribution is shown in Table 1.

The expected value E, of the attribute for label 1 is given in Eq. (1):

$$E = (a+b) * (a+c) /N \quad (1)$$

The observed values O of the attribute with categorical value1 for Label 1 is given as ' a' in Table 1. The chi-square value of the attribute for Label 1 is given in Eq. (2):

$$\chi^2 = (O-E)^2 /E \quad (2)$$

The chi-square values for all attributes are estimated and the corresponding weights are determined using the Eq. (3):

$$Weight\ of\ attribute_i = \chi_i^2 / \sum_{i=1}^{m} \chi_i^2 \quad (3)$$

A higher value of χ^2 shows a greater dependency of the attribute in discriminating the classes. A higher χ^2 value will result in greater weight implies the significance of the attribute (Rajalakshmi, 2014; Li *et al*., 2008). The weight values determined for the attributes using the above approach is used in the distance calculation during clustering. The distance d_{xy} between data objects x and y is computed using the Eq. (4):

$$d_{xy} = \sum_{i=1}^{m} w_i \, \delta x_i y_i \quad (4)$$

where,

$$\delta x_i y_i = \|x_i - y_i\|^2 \ if\ attribute_i\ is\ numeric$$
or $\delta x_i y_i = 0\ if\ x_i = y_i$
$\delta x_i y_i = 1\ if\ x_i \neq y_i$

m is the number of attributes, w_i is the relevance factor of the i^{th} attribute computed using the chi-square

Table 2: Contingency table for attribute type of chest pain

Chest pain type/class	Typical angina	Atypical angina	Non-anginal pain	Asymptomatic	Total
Absence	16	41	68	39	164
Presence	7	9	18	105	139
Total	23	50	86	144	303

Table 3: Contingency table for attribute age

Age/class	21-40	41-60	61-80	Total
Absence	12	117	35	164
Presence	6	89	44	139
Total	18	206	79	303

Table 4: Performance of mixed data clustering using the chi-square statistic distance

Data set	F-score	ARI	Accuracy
Heart disease	0.71	0.47	0.840
Credit	0.73	0.51	0.860
Vote	0.82	0.65	0.905
Bank	0.72	0.49	0.850

Table 5: Comparison of clustering accuracy using the chi-square statistic distance

Data set	Accuracy obtained using the proposed method	Accuracy reported in existing works
Heart disease	0.840	0.82 Ji *et al*. (2013)
Credit	0.860	0.77 Chatzis (2011)
Vote	0.905	0.86 He *et al*. (2011)

statistic approach using Eq. (3). This distance measure is used in the data clustering performed using k means algorithm. The clustering results obtained using this approach was comparatively better than the existing methods. The results are illustrated in the next section.

RESULTS AND DISCUSSION

The performance of the clustering algorithm using the proposed chi-square statistics based distance computation was observed for its improvement in clustering quality. The chi-square weighting scheme is illustrated for the Heart dataset for the attributes chest pain and age in Table 2 and 3. Using the values tabulated in Table 2 and 3 and Eq. (1) the observed and expected values O and E can be determined. The chi-square value and the corresponding attribute weight can be estimated using the Eq. (2) and (3).

The estimated chi-square values are 81.82 for chest pain attribute and 4.80 for age attribute and the corresponding attribute weights are 0.17 and 0.01. This approach is successful in identifying relevant attributes. The weights estimated using the proposed chi-square statistic approach is used in the distance computation between the data objects applying Eq. (4). The clustering when performed using this distance measure resulted in much better partitions and the clustering quality parameters such as accuracy, Adjusted Rand Index (ARI) and Fscore (Vendramin *et al*., 2009; Santos and Embrechts, 2009; Hubert and Phipps, 1985) of the obtained results is shown in Table 4.

Comparison with existing works: The accuracy measure of the clustering algorithm was compared with those existing works proposed by Ji *et al*. (2013), Ji *et al*. (2012), Chatzis (2011) and He *et al*. (2011) and shown in Table 5. It is evident that our proposed method is successful in producing better partitions. The results obtained for the UCI data sets are better than that reported in the literature. For the real time bank data set the results obtained using the existing approach and our method is given, for which also our method shows better performance.

CONCLUSION

In this research work we have proposed a chi-square statistic based approach to determine the distance between data objects defined by mixed numeric and categorical attributes. The proposed method is useful in determining the relevance of each of the attribute and is used as the weight factor in computing the distance during clustering. It is also successful in partitioning the data objects and exhibits better performance scores for ARI, Accuracy and F score. This study can be further extended to improve the computation of attribute relevance to be used as the corresponding weights in distance computation in mixed data clustering.

ACKNOWLEDGMENT

The authors would like to express their sincere thanks to the management of SSN College of Engineering, Chennai, India, for funding the High Performance Computing Lab (HPC Lab) where this research was carried out.

REFERENCES

Ahmad, A. and L. Dey, 2007. A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl. Eng., 63(2): 503-527.

Bai, L., J. Liang, C. Dang and F. Cao, 2011. A novel attribute weighting algorithm for clustering high-dimensional categorical data. Pattern Recogn., 44(12): 2843-2861.

Cao, F., J. Liang, D. Li, L. Bai and C. Dang, 2012. A dissimilarity measure for the k-Modes clustering algorithm. Knowl-Based Syst., 26: 120-127.

Cao, F., J. Liang, D. Li and X. Zhao, 2013. A weighting k-modes algorithm for subspace clustering of categorical data. Neurocomputing, 108: 23-30.

Chatzis, S.P., 2011. A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. Expert Syst. Appl., 38(7): 8684-8689.

Core Team, R., 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Han, J., M. Kamber and J. Pei, 2006. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

He, Z., X. Xu and S. Deng, 2011. Attribute value weighting in k-modes clustering. Expert Syst. Appl., 38(12): 15365-15369.

Hubert, L. and A. Phipps, 1985. Comparing partitions. J. Classif., 2: 193-218.

Ji, J., T. Bai, C. Zhou, C. Ma and Z. Wang, 2013. An improved k-prototypes clustering algorithm for mixed numeric and categorical data. Neurocomputing, 120: 590-596.

Ji, J., W. Pang, C. Zhou, X. Han and Z. Wang, 2012. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. Knowl-Based Syst., 30: 129-135.

Li, Y., C. Luo and S.M. Chung, 2008. Text clustering with feature selection by using statistical data. IEEE T. Knowl. Data En., 20(5): 641-652.

NIST, 2012. NIST/SEMATECH e-handbook of Statistical Methods. Retrieved form: http://www.itl.nist.gov/div898/handbook/2012.

Rajalakshmi, R., 2014. Supervised term weighting methods for URL classification. J. Comput. Sci., 10: 1969-1976.

Ralambondrainy, H., 1995. A conceptual version of the K-means algorithm. Pattern Recogn. Lett., 16(11): 1147-1157.

Santos, J.M. and M. Embrechts, 2009. On the use of the adjusted rand index as a metric for evaluating supervised classification. In: Alippi, C. *et al*. (Eds.), ICANN, 2009. Part 2 LNCS 5769, Springer, Berlin, Heidelberg, pp: 175-184.

UCI ML Dataset, year. UCI Machine Learning Repository. Retrieved form: http://archive.ics.uci.edu/ml. University of California, School of Information and Computer Sciences, Irvine.

Vendramin, L., Campello, R.J.G.B. Ricardo and E.R. Hruschka, 2009. On the comparison of relative clustering validity criteria. Proceeding of the 9th SIAM, International Conference on Data Mining (SDM). Sparks, NV, pp: 733-744.